Method and system for acquiring data from machine-readable documents

The present invention relates to a method and a system for acquiring data from machine-readable documents, the data being assigned to a database, in which individual data are extracted from the document as automatically as possible and are entered into corresponding database fields, the method and system according to the present invention relating to the acquisition of data in the case in which data cannot be extracted with the necessary degree of reliability for one or more particular database fields of a document.

Methods and systems for acquiring data from machine-readable documents are known. Standardly, the systems have a scanner with which documents are optically scanned. The data files produced in this way are machine-readable documents, and as a rule contain text elements. The text elements are converted into coded text with the aid of an OCR device. As a rule, predetermined forms or templates are assigned to the data files, so that on the basis of the forms data files containing particular items of information from the text can be determined in a targeted manner. These items of information are stored for example in a database.

Methods and systems of this sort are used for example in large firms in order to read invoices. The data extracted in this way can be communicated automatically to an accounting software program.

Such a system is described in US 4,933,979. This system has a scanner for the optical scanning of forms. In this system, a large number of types of forms can be defined, each type of form or template being defined by a plurality of parameters, in particular geometrically defined areas in which texts or images are to be contained. The form types can also be defined by additional characteristics, such as for example the type of script contained in the texts (letters, numbers, symbols, katakana, kanji, handwriting). After a form has been scanned, a template is assigned to the scanned form using a form type distinguishing device. Correspondingly, the data contained in the text field are read and extracted using an OCR device. If no suitable template exists, it is necessary to create one.

From WO 98/47098, another system is known for the automatic acquisition of data from machine-readable documents. Here, a scanner is used to optically scan forms. Subsequently, a line map of the form is created automatically. Here, on the one hand all lines are acquired, and all graphic elements are converted into a line structure. Other elements, such as for example text sections, are filtered out. All vertical lines form the basis for creating a vertical key, and all horizontal lines form the basis for creating a horizontal key. Subsequently, it is determined whether a template already exists having a corresponding vertical and horizontal key. If this is the case, the data are read out using a corresponding template. If this is not the case, then on the basis of the scanned-in form a template is created and stored using a self-learning mode.

In the book Modern Information Retrieval by Baeza-Yates and Ribeiro-Neto, Eddison-Wessley **[sic: Addison-Wesley]** Press, ISBN 0-201-39829-X, the basic principles of databases and information stored for rapid finding in databases are explained. Thus, in Chapter 8.2, a method using inverted data files, also designated an inverted index, is described. In this method, from a text that is to be examined first a dictionary is created having all the words contained in the text. Each word in the dictionary is assigned one or more numbers that indicate the location at which the word occurs in the text. Such inverted data files enable a more rapid automatic analysis of a text that is to be searched. In Chapter 8.6.1, a string matching method is described, in which two strings are compared and a cost measure is calculated that is indirectly proportional to the similarity of the strings. If the two strings are identical, the magnitude of the cost measure is zero. The more the strings differ, the greater is the magnitude of the cost measure. The cost measure is thus an expression of the similarity of the two strings. This and similar methods are also known under the names approximate string matching, Levenshtein method, elastic matching, and Viterbi algorithm. These methods are part of the field of dynamic programming.

In the not-yet-published patent application DE 103 42 594.2, a method and a system for acquiring data from a plurality of machine-readable documents are described in which, from a document that is to be processed -- the read document -- data are extracted by reading them out at positions in the read document that are determined by fields entered in a master document.

If an error occurs during the reading out of the read documents, the read document is displayed on a display screen and the data can be read out only by marking corresponding fields in the read document. Here, if it is required, additional master documents are automatically produced on the basis of the marked read documents, or existing master documents are correspondingly corrected. This system is easy enough to use that no special computer or software knowledge is necessary.

The present invention is based on the object of creating a method and a system for acquiring data from machine-readable documents in which the inputting of the data is significantly simplified in comparison with the known methods in cases in which data cannot be automatically extracted.

This object is achieved by a method having the features of Claim 1 and by a system having the feature of Claim 16. Advantageous constructions of the present invention are indicated in the respective subclaims.

With the methods explained above, data can be acquired from a plurality of machine-readable documents, the data being assigned to a database in that individual data are extracted from the document as automatically as possible and are entered into corresponding database fields. If data cannot be extracted with the necessary degree of reliability for one or more particular database fields of a document, for example because an error has been determined, caused for example by the fact that no data or false data are present in the document at the point at which the data are to be read, or that during the reading in of this document using an OCR method one or more characters are falsely converted, then according to the present invention the following steps are executed:

- displaying of the document on a display screen,
- indication on the display screen of the database field for which the data cannot be extracted with the necessary degree of reliability,
- execution of a proposal routine with which string sections in the vicinity of a pointer on the display screen that can be moved by a user are selected, marked, and proposed for extraction.

The document is displayed on the display screen so that the user can read it. In addition, the database field is indicated for which the data cannot be extracted with the necessary degree of reliability. In this way, the user is informed of the database field for which the data must still be extracted from the document shown on the display screen.

Through the execution or activation of the proposal routine, string sections in the vicinity of a pointer, movable on the display screen by the user, can be selected, marked, and proposed for extraction. In this way, the user need merely move the pointer on the document shown on the display screen into the vicinity of a string section that contains the data for the indicated database field. The data are then automatically selected, marked, and proposed for extraction. The user can then transfer **[or: incorporate]** the proposed string section into the database field merely by actuating a particular key.

Through the automatic selecting and marking of the string section, the process of incorporating the still-missing data is significantly simplified and accelerated.

According to a preferred specific embodiment of the present invention, during the selection of the string section concept **[or: conceptual, design]** information is taken into account that is assigned to the respective database field.

In the following, the present invention is explained in more detail in exemplary fashion on the basis of the drawing, in which:

Figure 1 shows a method for acquiring from a document data that cannot be extracted automatically,

Figures 2-6 each show copies of display screen representations corresponding to individual method steps of the method indicated in Figure 1,

Figure 7 shows a method for extracting data arranged in tables,

Figures 8, 9 each show a table with marked data, and

Figure 10 shows a system for executing the method according to the present invention.

The method according to the present invention for acquiring data from machine-readable documents is a development of the methods described above with which data can be extracted from documents and stored in a database by machine.

However, in these methods it is not always possible to fill all database fields of the database reliably with data extracted from the documents. If, for example, there is an error during the extraction of the data, the automatic method is interrupted and, with the cooperation of the user, the data from the document are manually entered into database fields. Such an error can result from the fact that in the document to be processed no suitable string section is found from which the data can be read, or the string section contains errored data that arise for example during the conversion of the document into coded text using an OCR method.

The method according to the present invention thus begins when data cannot be reliably extracted. The expression "not reliably extractable" includes both fundamental errors in the reading of data that make a reading of the data impossible, and also read data that are mapped to the database field while taking into account context information, the quality of the mapping being determined during this process. Such mapping methods include for example the string matching method named above. If the mapping quality achieved here is too low, the automatically read-in data are evaluated as insufficiently reliable and are rejected.

In the following, the method according to the present invention is explained on the basis of the flow diagram shown in Figure 1. In the flow diagram, all steps that are executed automatically are identified with an "a" in a circle, and all steps that are to be carried out manually by the user are identified with a "m" in a circle.

The method begins with step S1.

When data for at least one database field cannot be extracted with the necessary degree of reliability, the corresponding document 1 is displayed on a display screen 2, and the database field 3 is indicated (step S2). Figure 2 shows a display screen representations immediately after the determination that data could not be extracted with the necessary degree of reliability; here the document 1 is shown in a window 4/1 on the right side of the display screen representation. Two windows 4/2 and 4/3 are situated on the left side. Window 4/2 contains an overview of the documents that are to be processed, and in window 4/3 the individual database fields are indicated in which data are stored that are to be read from document 1.

In the example shown, none of the database fields could be filled with data, for which reason the individual database fields 3 are provided with the designation [empty]. However, it is also possible for data to be missing only in a few database fields, or only in a single database field.

In Figure 2, the database field "InvoiceNumber" is marked darker in comparison to the other database fields 3, which indicates to the user that data are to be extracted from document 1 for this database field 3. In addition, in the upper area of window 4/1 the term "InvoiceNumber" is indicated in a larger font, additionally indicating to the user the database field for which data are to be extracted.

In window 4/1, the user can now position a pointer 5 that he preferably situates in such a way that it is located as close as possible to the string section for which the user assumes that the content is to be stored in the corresponding database field. In the example shown in Figure 2, data are to be extracted for the database field "InvoiceNumber," so pointer 5 is positioned in the vicinity of invoice number "4361" (step S3).

Here, pointer 5 can be moved in window 4/1 using a mouse 6 or via inputs on a keyboard 7.

After the positioning of pointer 5, a proposal routine begins that comprises a plurality of method steps. This proposal routine can on the one hand be initiated in that pointer 5 is not moved for a predetermined time interval, whereupon the proposal routine is then

automatically executed, or it can be initiated by actuating a particular mouse button or keyboard key.

In step S4, it is first checked whether there is located in the immediate vicinity of the pointer a string section having a concept suitable for database field 3, insofar as concept information has been previously assigned to the corresponding type of the database field. This concept information includes the syntax and/or the semantics of the database field. Information concerning syntax includes for example the number of numerals and/or letters and/or specified formats of the string section that is to be read. Thus, date fields, amount fields, and address fields have as a rule particular formats. Semantic information includes specified terms that can be entered into the corresponding database field. This is useful for example for currency indications, or if the article designation of a particular supplier that can supply a limited number of articles is to be read in. The corresponding article designations are then stored in a lexicon and can then be unambiguously recognized.

In the exemplary embodiment shown in Figure 2, the two string sections "4361" and "02.08.2002" are situated in the vicinity of pointer 5. The latter string section has the syntax of a date, and for this reason it is rejected for the extraction of the invoice number. The string section "4361" corresponds to the syntax of an invoice number. Therefore, in step S4 it is decided that a string section having a suitable concept is present, and for this reason the method sequence next goes to step S5. In step S5, the string section "4361" is marked (Figure 3). In the present exemplary embodiment, the marking takes place through a colored highlighting **[or: background]** of the string section and through the drawing in of a frame 8.

If in step S4 no suitable concept is determined, the method sequence goes to step S6. In step S6, the individual character situated closest to pointer 5 is determined, which, in the present exemplary embodiment according to Figures 2-4, is the "1." Subsequently, the boundaries of the string section containing this character are determined according to general rules. These boundaries can for example be determined by empty characters or empty spaces in the document 1, or by particular punctuation marks or other markings in document 1. If corresponding boundary markings are recognized, the string section situated between them is selected and marked. In the exemplary embodiment shown in Figures 2 and 3, on each side of

string section "4361" there are situated empty spaces, via which an unambiguous selection of the marking of the string section is possible, according to the general rules as well.

Independent of whether the string section has been selected or marked according to step S5 or according to step S6, the method sequence goes to method step S7, with which the string section is displayed in an additional frame 9 as a coded text, and is displayed in an enlarged fashion in another frame 10 (Figures 3, 4). In the present exemplary embodiment, document 1 is present as a graphic data file, e.g. in the .pdf, .tif, .gif, or .jpg format. Standardly, in the preceding method segment the document was subjected to an OCR routine and converted into coded text. The coded text is here also examined for concepts, and the corresponding information is stored. The section corresponding to the string section is removed from this coded text and is shown in frame 9. In this way, the user recognizes whether the string section has been correctly converted into coded text.

In frame 10, the string section is shown in a graphic format in an enlarged representation, so that the user can also recognize details in the string section.

In step S7, the proposal routine is terminated.

In step S8, the user judges whether the selected and marked string section is fundamentally suitable for transferring into the database field. If this is not the case, pointer 5 is repositioned (S3) and the proposal routine (S4 – S7) is executed again. If, in contrast, the selection of the string section is fundamentally suitable, the user judges whether the marked area is also correct (step S9). If this is not the case, the user can manually process the marking of the string section and/or can edit the coded text in frame 9 (step S10). With the editing of the coded text, errors resulting from an incorrect OCR conversion can be removed. When these corrections (adapt area, edit) are made, the marked area and the contents of frames 9 and 10 are automatically adapted.

If the marked area is correct or has been correspondingly revised by the user, the method sequence moves to step S11, in which the data contained in the selected string section are transferred into the corresponding database field (Figure 4). This transferring of the data is

initiated by user actuation of a predetermined mouse button or key on the keyboard. Subsequently, the method for extracting data for a database field is terminated (S12). If data are to be read for additional database fields, the method begins again with step S1. In Figure 5, the next database field to be read ("Invoice Date") is indicated.

With the method according to the present invention, the activity of a user in the manual transferring of data from a document into a database field is limited to the positioning of the pointer, the checking of the automatically proposed selection and the possible correction of the area, and the actuation of a key in order to transfer the data. The selection and the marking of the area of the string section to be selected are carried out automatically by the method according to the present invention.

Figures 2 to 5 show the transfer of data into an individual database field. However, by taking into account concept information, it is also possible to extract data for a plurality of database fields with a single string section. Figure 6 shows a corresponding exemplary embodiment, in which the complete address is marked and read as a string section, the address being automatically segmented into the individual database fields name, company, street, postal code, and city.

In the following, another construction of the method described above, with which data can be extracted from tables, is explained on the basis of the flow diagram from Figure 7 and the display screen representations according to Figures 8 and 9.

This method begins with step S15.

In step S16, the values of the table in the first table row are extracted according to the above method through the positioning of the pointer, the automatic selection and marking of the string section, and the transferring of the data into corresponding database fields. Figure 8 shows a table in which the string sections of the first table row are marked that have been transferred into the corresponding database fields. These database fields have the structure of a table; for example, they are applied as a two-dimensional data field, so that during the

extraction of the data into these database fields the method recognizes automatically, on the basis of the database field, that data are being read out from a table.

A row of a table can also extend over a plurality of pages if the table correspondingly extends over a plurality of pages. If the data of the first table row has been completely extracted, the user can initiate the automatic extraction of the further table entries using a predetermined input. If this input is actuated by the user, then, in step S17, first a list is created of all string sections that are situated under the first table row.

In step S18, a cost function is used to determine a cost value between sequences of string sections of the list and the sequence of the string sections of the first table row, on the basis of which data were extracted into the database fields in step S16. In this cost function, low costs are assigned to the sequences of the string sections of the list whose string sections agree with, or are at least very similar to, the corresponding string sections of the first table row, with respect to their horizontal position and their width. This cost value is thus indirectly proportional to the degree of similarity between the sequences of string sections appearing in the list and the sequence of string sections contained in the first table row.

The cost function used here corresponds to the cost function described in Chapter 8.6.1 of String Matching Allowing Errors in Modern Information Retrieval (ISBN 0-201-39829-X), with which an individual cost value between a string section of the first table row and a string section of the further table rows is determined. Because each sequence comprises a plurality of string sections, the Viterbi algorithm is used to calculate an overall cost value or overall similarity value for each of the individual sequences of string sections, through summation of the individual cost values.

On the basis of these cost values or similarity values, the sequences of string sections are determined as table rows whose similarity value lies beneath a predetermined threshold value (S19). In this way, all table rows, and thus table entries, of the table are determined. They are marked in step S20 (Figure 9) and in step S21 they are extracted, i.e., automatically read out, converted into coded text if necessary, and stored in the corresponding database fields.

In step S22, this method is terminated.

Usefully, it is possible to post-process the table entries, i.e., to modify (move, enlarge, make smaller) the marked areas, or to remove or add individual rows. In the case of a post-processing, the entries in the database fields are automatically updated correspondingly.

In addition, during reading out of the data and entering into the database fields an additional check can take place through a mapping using the string matching method, with which it is determined how well the entries agree with the concept specified by the individual database fields.

In addition, the method according to the present invention can be combined with the method described in German patent application DE 103 42 594.2 for acquiring data from a plurality of machine-readable documents, for which reason reference is made to the complete content of this patent application, and it is incorporated into the present patent application by reference.

In this method for the automatic acquisition of data from a plurality of machine-readable documents, master documents are compared with a read document and their similarity is evaluated. The method applied here can also be used for reading out from a table, the sequence of the selected string sections corresponding to the first table row of the master document, and the combinations of string sections corresponding to the further table rows of the read documents.

In the above-described method according to the present invention for extracting data from tables, a user need merely move the pointer to the table entries in the first table row and confirm the transferring of the then automatically selected and marked string sections as data for the corresponding database field. After the user has done this for all table entries of the first table row, he need merely initiate the complete reading out of the further table entries by making an input. The method then automatically determines the further table entries, marks them, and extracts the data into the database.

This significantly accelerates the reading out of data from the table into a database. Method segment S17 to S21 therefore represents an independent invention in its own right, which is however preferably applied in combination with the method represented in Figure 1, to which step S16 relates.

Figure 10 schematically shows a system for executing the method according to the present invention. This system 11 comprises a computer 12 having a storage device 13, having a central processor device (CPU) 14, and having an interface device 15. A scanner 16, a display screen 2, and an input device 17 are connected to computer 12. Input device 17 includes a keyboard 7 and/or a mouse 6.

In storage device 13, a software product is stored for executing the method according to the present invention, this software product being executed at CPU 14. Scanner 16 is used to acquire documents and to convert them into an electronic data file. These electronic data files are read by computer 12 and are preprocessed if necessary, using an OCR routine and/or a method for recognizing particular syntax or semantics in the data file. Subsequently, the documents contained in the data files are processed in a manner corresponding to the method described above, using system 11. At input device 17, the corresponding inputs can be carried out, these being limited to movements of pointer 5 and a few keyboard inputs. If necessary, the marked fields can be moved using the keyboard or the mouse, or can be adapted by enlargement or by being made smaller, or the coded text can be edited.

The present invention has been explained above on the basis of an exemplary embodiment. Modifications thereof are possible within the scope of the present invention. Thus, for example, instead of frame 8 it is possible to provide only frame 10, in which the selected string section is shown in an enlarged manner. This frame 10 also represents a marking of the string section.

In the above-explained exemplary embodiment, the documents are scanned in and are then present in a graphic format. However, the method according to the present invention can also be used for reading information from documents that are already present in coded text, such

as for example e-mails. Of course, given such an application it is not necessary for the documents to be converted into coded text using an OCR routine.

Consequently, the present invention can be briefly summarized as follows:

The present invention relates to a method for acquiring data from machine-readable documents, the data being assigned to a database.

With the present invention, string sections located in the vicinity of a pointer that can be moved by the user are automatically selected and marked, and their content is proposed for transfer into a database.

According to a development of the method according to the present invention, the content of a table can be read out in a fully automatic manner if the table entries in a first table row have already been read out according to the above method.

Exemplary embodiments of the present invention have been described. Here it is clear that someone skilled in the art can at any time indicate modifications and developments that make use of the concept of the present invention. In addition, the present invention can be realized both by means of electronic components (hardware) and through computer program elements (software or software modules). In particular, the present invention is realized here as a combination of electronic hardware elements and software elements. Correspondingly, the present invention also includes computer program products, such as for example electronic data carriers (CDs, DVDs, diskettes, tape drives), or components that are distributed via computer networks (Internet) and/or on computers, and in particular are loaded into intermediate storage units and are kept ready there and/or are run from there.

List of reference characters

1	document
2	display screen
3	database field
4	window
5	pointer
6	mouse
7	keyboard
8	frame
9	frame
10	frame
11	system
12	computer
13	storage device
14	CPU
15	interface device
16	scanner
17	input device